

5                   **SYSTEM AND METHOD FOR PROVIDING DEFINITIONS**

**Cross-Reference to Related Application**

                    This non-provisional patent application claims priority under 35 USC § 119(e) to U.S. provisional patent application, Serial No. 60/472,445, filed May 20, 2003, the disclosure of which is incorporated by reference.

10                   **Field of the Invention**

                    The present invention relates in general to providing definitions and, in particular, to a system and method for providing definitions.

**Background of the Invention**

                    A system and method for providing definitions is described. There is a  
15                   vast amount of content available on the Internet. Some of this content is organized in the form of glossaries or definitions. The system and methods described herein allow one to tap into these available resources to quickly and efficiently provide definitions for phrases. "Phrases" may refer to words, phrases, or any other semantic unit that is capable of definition.

20                   **Summary of the Invention**

                    An embodiment provides a system and method for providing definitions. A phrase to be defined is received. One or more documents, which each contain at least one definition, are determined. The phrase is matched to at least one of the definitions. One or more definitions for the phrase are presented.

25                   A further embodiment provides determining definitions from distributed information stores. One or more documents are identified. Each document is maintained in a distributed information store and contains a definition for an associated phrase. Information regarding each identified document is stored. A phrase for which a definition is sought is matched against the stored information

for each identified document. Each identified document is fetched from the distributed information store and one or more matching definitions are returned. Each matching definitions is presented.

5 Still other embodiments of the present invention will become readily apparent to those skilled in the art from the following detailed description, wherein are described embodiments of the invention by way of illustrating the best mode contemplated for carrying out the invention. As will be realized, the invention is capable of other and different embodiments and its several details are capable of modifications in various obvious respects, all without departing from  
10 the spirit and the scope of the present invention. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature and not as restrictive.

### **Brief Description of the Drawings**

The patent or application file contains at least one drawing executed in  
15 color. Copies of this patent or patent application publication with the color drawings will be provided by the Office upon request and payment of the necessary fee.

FIGURE 1 is a block diagram showing a system for providing definitions, in accordance with the present invention.

20 FIGURE 2 is a block diagram showing a computer system for use in the system of FIGURE 1.

FIGURE 3 is a flow diagram showing a method for providing definitions, in accordance with the present invention.

25 FIGURE 4 is a screen shot showing, by way of example, definitions provided by the method of FIGURE 3.

FIGURE 5 is a screen shot showing, by way of example, further definitions provided by the method of FIGURE 3.

FIGURE 6 is a screen shot showing, by way of example, still further definitions provided by the method of FIGURE 3.

### **Detailed Description**

30

## System Overview

FIGURE 1 is a block diagram showing a system 10 for providing definitions, in accordance with the present invention. A plurality of individual clients 12 are communicatively interfaced to a server 11 via an internetwork 13, such as the Internet, or other form of communications network, as would be recognized by one skilled in the art. The individual clients 12 are operated by users 19 who transact requests for Web content and other operations through their respective client 12.

In general, each client 12 can be any form of computing platform connectable to a network, such as the internetwork 13, and capable of interacting with application programs. Exemplary examples of individual clients include, without limitation, personal computers, digital assistances, "smart" cellular telephones and pagers, lightweight clients, workstations, "dumb" terminals interfaced to an application server, and various arrangements and configurations thereof, as would be recognized by one skilled in the art. The internetwork 13 includes various topologies, configurations, and arrangements of network interconnectivity components arranged to interoperatively couple with enterprise, wide area and local area networks and include, without limitation, conventionally wired, wireless, satellite, optical, and equivalent network technologies, as would be recognized by one skilled in the art.

For Web content exchange and, in particular, to transact searches, each client 12 executes a Web browser 18 ("Web browser"), which implements a graphical user interface and through which search queries are sent to a Web server 20 executing on the server 11, as further described below with reference to FIGURE 2. Each search query describes or identifies information, generally in the form of Web content, which is potentially retrievable via the Web server 20. In addition, the search query can include a phrase for which a definition is sought, as further described below with reference to FIGURE 3. The search query provides characteristics, typically expressed as terms, such as keywords and the like, and attributes, such as language, character encoding and so forth, which enables a search engine 21, also executing on the server 11, to identify and send

back Web pages. The terms and attributes are a form of metadata, which constitute data describing data. Other styles, forms or definitions of search queries, search query characteristics, and metadata are feasible, as would be recognized by one skilled in the art.

5           The Web pages are sent back to the Web browser 18 for presentation, usually in the form of Web content titles, hyperlinks, and other descriptive information, such as snippets of text taken from the Web pages. The user can view or access the Web pages on the graphical user interface and can input selections and responses in the form of typed text, clicks, or both. The server 11  
10 maintains an attached storage device 15 in which Web content 22 is maintained. The Web content 22 could also be maintained remotely on other Web servers (not shown) interconnected either directly or indirectly via the internetwork 13 and which are preferably accessible by each client 12.

          The search engine 21 preferably identifies the Web content 22 best  
15 matching the search query terms to provide high quality Web pages, such as described in S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Search Engine" (1998) and in U.S. Patent No. 6,285,999, issued September 4, 2001 to Page, the disclosures of which are incorporated by reference. In identifying matching Web content 22, the search engine 21 operates on  
20 information characteristics describing potentially retrievable Web content, as further described below with reference to FIGURE 2. Note the functionality provided by the server 20, including the Web server 20 and search engine 21, could be provided by a loosely- or tightly-coupled distributed or parallelized computing configuration, in addition to a uniprocessing environment.

25           The individual computer systems, including server 11 and clients 12, include general purpose, programmed digital computing devices consisting of a central processing unit (processors 13 and 16, respectively), random access memory (memories 14 and 17, respectively), non-volatile secondary storage 15, such as a hard drive or CD ROM drive, network or wireless interfaces, and  
30 peripheral devices, including user interfacing means, such as a keyboard and display. Program code, including software programs, and data is loaded into the

RAM for execution and processing by the CPU and results are generated for display, output, transmittal, or storage. The Web browser 18 is an HTTP-compatible Web browser, such as the Internet Explorer, licensed by Microsoft Corporation, Redmond, WA; Navigator, licensed by Netscape Corporation, Mountain View, CA; or a Mozilla or JavaScript enabled browser, as are known in the art.

#### Computer System Components

FIGURE 2 is a block diagram showing a computer system 30 for use in the system 10 of FIGURE 1. The computer system 30 includes a processor 31 and visual display 32, such as a computer monitor or liquid crystal diode (LCD) display, as are known in the art. The computer system 30 executes a Web browser 18 (shown in FIGURE 1), which implements a graphical user interface 37. Visual Web content, including retrieved definitions, is output within a display area defined on the graphical user interface 37 while user inputs are generally input both within the display area and within specified user input regions. Textual user inputs are received via a keyboard 33. Linear, non-textual inputs are received via a pointing device 34, such as a mouse, trackball, track pad, or arrow keys. Similarly, voice- and sound-based inputs are received via a microphone 35. Visual outputs are displayed via the graphical user interface 37 on the visual display 32, while audio outputs are played on the speakers 36. Other forms of computer components, including processor 31, visual display 32, and input and output devices could be used, as would be recognized by one skilled in the art.

#### Method Overview

One embodiment of the present invention will now be described with reference to FIGURE 3, which provides a flow diagram showing a method for providing definitions, in accordance with the present invention. The method is described as a sequence of process operations or steps, which can be executed, for instance, by the system of FIGURE 1, or equivalent component.

First, a phrase for which definition is sought is provided (block 310). The phrase may be provided by, for example, a user request or query, or by any other

means. One example of a system for providing a phrase is that located at the URL identified by *http://labs.google.com/glossary*, the contents of which are incorporated by reference. In addition, the spelling of the phrase can be corrected if necessary or normalized into a common root form to provide more consistent definition results.

Documents that contain definitions are determined (block 320). These documents may be determined in any number of ways. For example, such documents may be determined during Web-crawling or spidering performed by search engines in either real time or batch processing modes. Once a document is determined to contain definitions, the document (or information about the document, such as the document's URL) may be stored or remembered for future use. "Authoritative" sources for definitions may also be used, for example, documents associated with Web sites, such as *http://www.dictionary.com*.

In one embodiment of the present invention, documents containing definitions are located substantially in real time, by conducting a query via an Internet search engine. In a further embodiment, the documents are located substantially in a batch processing mode, for example, by fetching, parsing and indexing the documents containing definitions off-line prior to receiving queries. In addition, a combination could be used, such as by providing batch processing for identifying documents containing definitions and using real time processing to fetch, de-duplicate and clean up definitions responsive to each query.

The query may search for terms that are likely to indicate the presence of definitions, such as "glossary," "definition," "dictionary," and so forth, as well as variants and canonicalizations thereof. The search may be conducted over the document text as a whole, or may be restricted to certain portions or fields within documents, such as the title field, fields containing other metadata, and so forth. The structure of documents, that is, the tagged nature of HTML documents, may also be relevant to determining how to structure the query. In an embodiment of the invention, a search for "glossary," "definitions," or "dictionary" in the title of Web pages are used to determine the relevant documents. As will be recognized by one of ordinary skill in the art of information retrieval, the above methods may

be combined in various fashions and with numerous other methods to determine definition containing documents.

The phrase for which definition is sought is then matched against the determined documents to return definitions (block 330). The documents determined in this step (block 330) may be parsed to identify occurrences of the phrase being sought and the phrase's associated definition. For example, definition containing documents may be organized with "headwords," or words that can be looked up in a dictionary form. There are various methods for identifying headwords and/or identifying definitions. In one embodiment of the invention, one or more of the following methods are used to parse apart documents, identify headwords, and/or return definitions:

- If the page uses `<dl>`, `<dt>` and `<dd>`, which are HTML tags used for specifying lists of definitions, the HTML mark up is relied upon to identify definitions, that is:

```
An example definition list
<dl>
  <dt>Headword 1
    <dd>This is the definition of Headword 1
  <dt>Headword 2
    <dd>This is the definition of Headword 2
  <dt>Headword 3
    <dd>This is the definition of Headword 3
</dl>
```

- HTML tags, such as `<p>`, `<tr>`, `<li>`, and `<br>`, may be treated as separators between successive definitions.
- White space or punctuation (., : -) is eliminated at the beginning of definitions.
- Headwords may be identified by the fact that the headwords are surrounded by the HTML tags `<b>`, `<strong>`, `<em>`, `<code>`, or `<span>`.
- Lines that do not start with headwords are deleted.
- If there are fewer than N, for instance, N = 5, definitions found in the document or page, all definitions in the document or page are discarded.

The parser does not need to be perfect at identifying all headwords and definitions. In one embodiment, due to the large number of definition-containing documents determined in the definition document determination step (block 320), the parser is biased towards precision rather than thoroughness. In other words, the parser errs towards throwing entries away rather than keeping entries that may be incorrect because there are more than enough definitions to supply a satisfactory outcome. Similarly, in a further embodiment, the parser de-duplicates entries that are duplicative or merely cumulative of other entries.

One or more of the returned definitions are then provided (block 340). In one embodiment, the returned definitions are ranked according to PageRank™ of the documents from which they are retrieved, according to the methods disclosed in U.S. Patent No. 6,285,999, cited above. The retrieved definitions may also be processed for presentation, such as by carrying out one or more of the following steps:

- Removing:
  - all HTML markup;
  - leading and trailing white space in both headword and definition;
  - all punctuation: ( . : ; ! ? - ) in the headword;
  - all leading non-alpha and non-parenthesis in the headword and definition;
  - all trailing non-alphanumeric and non-parenthesis in the headword.
- Throw the definition away if:
  - the definition starts with “see”
  - the definition is a duplicate of one already retrieved.
- Capitalize the first letter the definition.

In one embodiment, only definitions whose head phrases are an exact match for the phrase are presented. However, in other embodiments of the invention, a looser form of matching may be allowed.

Other information may also be determined and presented. In one embodiment of the present invention, superstrings of the phrase are tabulated and presented as query refinements or related phrases. Superstrings are strings that



contain the phrase (or possibly common variants or canonicalized versions of the phrase). Methods for determining common variants or canonicalized versions of words and phrases are described in, for example, U.S. patent application Serial No. 10/377,117, Attorney Docket No. GP-091-00-US, entitled "SEARCH  
5 QUERIES IMPROVED BASED ON QUERY SEMANTIC INFORMATION,"  
filed March 3, 2003, pending, and listing Amit Singhal et al. as inventors, which disclosure is incorporated by reference. For example, the top M superstrings may be listed. Similarly, the phrase may be presented in a processed form, such as in the phrase's most common capitalization; for instance, a user query for [pocket  
10 pc] or [pocket pcs] may be presented as "Pocket PC" because that is the most common form and/or capitalization found in the definitions.

As will be recognized by one of skill in the art, the steps described above with reference to FIGURE 3 need not be performed in the order listed, and steps may be added or removed.

15 As used in this specification, a "document" is to be broadly interpreted to include any machine readable or machine storable work product. A document may be a file, a combination of files, one or more files with embedded links to other files, and so forth. The files may be of any type, such as text, audio, image, video, and so forth. In the context of the Internet, a common document is a Web  
20 page, as is known in the art.

According to a further aspect of the invention, in situations where no definitions are found (or where definitions are not selected for presentation, such as if there is doubt as to whether the definition properly matches the original provided phrase), a set of terms or phrases that are related to the original phrase,  
25 that are deemed likely to be related to the phrase, that may be of interest (e.g. of interest to the user entering the original phrase), or even a "random" or eclectic set of terms or phrases for which definitions are returned, may be provided. Such terms may be provided, for example, to give a user a guide as to the types of terms that are defined, or for user amusement.

30 Sample Web Pages

FIGURE 4 is a screen shot 400 showing, by way of example, definitions provided by the method of FIGURE 3. A glossary search for the phrase “rdbms” is provided, substantially as shown.

5      FIGURE 5 is a screen shot 500 showing, by way of example, further definitions provided by the method of FIGURE 3. A glossary search for the phrase “pocket pc” is provided, substantially as shown.

FIGURE 6 is a screen shot 600 showing, by way of example, still further definitions provided by the method of FIGURE 3. A glossary search for the phrase “pocket pcs” is provided, substantially as shown.

10      While the invention has been particularly shown and described as referenced to the embodiments thereof, those skilled in the art will understand that the foregoing and other changes in form and detail may be made therein without departing from the spirit and scope of the invention.